DOCUMENT RESUME

ED 104 940 TM 004 398

AUTHOR Jaeger, Richard M.
TITLE Some Psychometric Indicators for Statewide Assessments.
PUB DATE [Apr 75]
NOTE 61p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D. C., March 30-April 3, 1975). Document not available in hard copy due to marginal legibility of original document

EDRS PRICE MF-$0.76 HC Not Available from EDRS..PLUS POSTAGE
DESCRIPTORS Cultural Factors; *Educational Assessment; Institutions; *Measurement Techniques; Objectives; Sampling; Standard Error of Measurement; *State Programs; Taxonomy; Test Bias; *Test Reliability; *Test Validity

ABSTRACT

Three new indicators of psychometric quality for objectives-based statewide assessments are proposed. These measures provide indication of the stability of reported data on item and objectives mastery, the validity of assessment items for members of various cultural groups, and the convergent validity of prescribed objectives mastery scores. The results provided should also have application in situations other than statewide assessments. In particular, the results should be applicable whenever the psychometric quality of measurements for institutions, rather than individuals, is of concern. (Author/RC)

SOME PSYCHOMETRIC INDICATORS FOR STATEWIDE ASSESSMENTS[1]

by

Richard M. Jaeger
University of South Florida

Prolegomenon

Strictly speaking, this paper is true to its title: it does provide some psychometric indicators for statewide assessments. However its distribution of content is skewed, in that indices of stability are given far greater attention than are indices of validity. The results provided should also have application in situations other than statewide assessments. In particular, the results should be applicable whenever the psychometric quality of measurements for institutions, rather than individual , is of concern.

A review of state accountability legislation reveals that states allege a multiplicity of purposes for their assessments (Hawthorne, 1974). Some legislatures mandate uniform measurement of all pupils, presumably to provide bases for individual decisions. More often, however, the legislative objectives of assessment require aggregated information on pupils in various institutions--schools, school districts, or specific educational programs. The results of these measurements -e intended to provide bases for decisions concerning the institutions, rather than the individual pupils they serve.

Examples of legislation that motivate institutional measurement include the Connecticut State Legislature's Public Act Number (1971), that requires the State Board of Education to develop an assessment procedure to measure the adequacy and effectiveness of educational programs in Connecticut's public schools; Georgia State Senate Bill Number 672 (1974), that requires the State Board of Education

[1]Presented as part of a symposium on Advances in the Methodology of Statewide Assessment, at the Annual Meeting of the American Educational Research Association, Washington, D.C., April, 1975.

to establish performance-based criteria to evaluate the instructional program of each school in the state; and Section 290.1 of the Pennsylvania School District Reorganization Act of 1963, that requires the State Board of Education to develop an evaluation procedure for objectively measuring the state's educational programs.

With statewide educational assessment has come increased attention to techniques and procedures for measurement. Altnough E. L. Thorndike defined the difference between criterion-referenced and norm-referenced measurement in 1918, and the Boston Public Schools conducted a criterion-referenced assessment in 1916, both the term and the practice are enjoying a renaissance that would make one doubt their earlier origins. The current measurement literature abounds with statements on the relative worth of criterion-referenced and norm-referenced measurement, and articles on methods for assessing the reliability and validity of criterion-referenced measures are numerous (Stanley, 1971; Livingston, 1972a, 1972b, 1973; Harris, 1972, 1973; Ebel, 1973; Popham and Husek, 1969).

Since many statewide assessment programs attempt criterion-referenced interpretations of their measurements, one might think that psychometric indices for such measurements would be sufficient for state assessments. They are not. The very concept of reliability, although inherently generic, has been developed in the context of measuring individuals (Lord and Novick, 1968, p. 61; Stanley, 1971, p. 357; Cronbach, 1951). New reliability formulations intended for use with criterion-referenced measures (Livingston, 1972a) have also been proposed as indices of the stability of individual assessments. Indices of stability for institutional measures are therefore still to be developed. One could support a similar case for indices of validity.

Indicators of Stability

The classical definition of reliability. Lord and Novick (1968, p. 61) define the reliability of a test as "a measure of the degree of true-score variation relative to observed-score variation." As is typical, they refer to variation among the true scores and observed scores of individuals. The Lord and Novick definition of reliability can be adapted directly to an index of stability for the mean measurement performance of a group of pupils. The formulation is as follows:

Consistent with classical test theory, assume that the observed measurement for the i-th person is composed of true-score and error terms,

$$X_i = T_i + E_i;$$

that errors are uncorrelated with true scores, and that errors are uncorrelated across persons;

$$\rho_{T_i,E_i} = 0,$$

$$\rho_{E_i,E_j} = 0, \quad i \neq j.$$

Assume in addition, that the mean is based upon measures of n persons, chosen from some larger population through simple random sampling (Cochran, 1963).

Given these assumptions, the reliability of an individual's score is equal to

$$\rho_{XX} = \sigma^2_{T_i}/\sigma^2_{X_i} = \sigma^2_{T_i}/\sigma^2_{T_i+E_i} = \sigma^2_{T_i}/(\sigma^2_{T_i}+\sigma^2_{E_i})$$

and the "reliability" of the group mean score equals

$$\rho_{XX} = \frac{\sigma^2_{\overline{T}}}{\sigma^2_{\overline{X}}} = \frac{\sigma^2_{\overline{T}}}{\sigma^2_{\overline{T}+\overline{E}}} = \frac{\sigma^2_{\overline{T}}}{\sigma^2_{\overline{T}} + \sigma^2_{\overline{E}}} = \frac{\sigma^2_{T/n}}{\sigma^2_{T/n} + \sigma^2_{E/n}} = \frac{\sigma^2_{T}}{\sigma^2_{T} + \sigma^2_{E}} = \rho_{XX},$$

where $\overline{X} = \frac{\Sigma X_i}{n} = \frac{\Sigma(T_i + E_i)}{n} = \overline{T} + \overline{E}.$

As traditionally defined, then, the reliability of the mean score of a randomly sampled group of persons is identical to the reliability of the score of an individual sampled from the same population.

Although it may be comforting to be in familiar territory with the reliability of group mean scores, the interpretation of the index is not clear. If, for example, $\overline{X}$ represents the mean achievement test score of the fourth-graders in a single school, one might attempt test-retest estimation of the reliability of the mean by randomly sampling schools, administering the test on successive occasions to all fourth-graders in each sampled school, and computing the correlation between individuals' successive scores. If the sampled schools could be considered representative of schools in some larger administrative unit, or perhaps in the nation, and if fourth-graders could be assumed to be randomly allocated among schools, the resulting reliability coefficient would be an estimate of the stability of school means. The magnitude of the corresponding standard error of measurement could be evaluated by considering national norms for school mean achievement scores (U.S. Government Printing Office, 1974). To avoid the assumption of random allocation of pupils to schools, one could compute the ecological correlation of school means for successive administrations of the test. However, the population value of the resulting coefficient would not then equal the reliability coefficient for individuals. An estimate of reliability consistent with the Lord and Novick definition would result only if true-score means and error means were uncorrelated across schools, and mean errors resulting from successive administrations of the test were uncorrelated across schools. Since some school environmental factors would likely contribute to error in consistent ways, it is doubtful that the latter assumption could be met.

A Generalization -- Universes and Universe Scores. Rather than defining the stability of group mean scores through direct extrapolation of classical test theory, it would seem more productive to first give attention to the meaning desired in such an index. The theoretical underprinnings for the generalization proposed here are contained in Cronbach, Gleser, Nanda and Rajaratnum (1972). A specific citation provides the needed background:

"A behavioral measurement is a sample from the collection of measurements that might have been made, and interest attaches to the obtained score only because it is representative of the whole collection. If the decision maker could, he would measure the person exhaustively and take the average over all measurements."

"Educators and psychologists have traditionally referred to the average reached via exhustive measurement as 'the true score' for the person. We speak instead of a universe score. This emphasizes that the investigator is making an inference from a sample of observed data, and also that there is more than one universe to which he might generalize. Any person fits within many defferent populations. . . . Any observation fits within a variety of universes."

If references to persons are replaced with references to institutions, the major concepts in the paragraphs cited above still apply. Any measurement on an institution is a sample from a population of measurements that might be made. And the institution(s) measured constitute a sample from several potential populations of institutions. An index of the stability of measurement of an institution should specifically reflect the generalizations desired, both with respect to a population of potential measurements, and to a population of institutions. Indices of stability referenced to specific generalizations will facilitate unambiguous interpretations.

A Taxonomy of Universe Scores for Statewide Assessments. Since the objectives of statewide assessment vary among the states and a given statewide assessment may have several purposes, it is not surprising that a number of different universe scores could be of interest. At least five dimensions can be used to structure a taxonomy of universe scores and corresponding estimators for statewide assessments. These dimensions are

1) The evaluative referent for interpretation of the universe score:
   a) domain of content or abilities
   b) normative
2) The type of statistic that constitutes the observed score:
   a) the proportion of examinees that answers a question correctly
   b) the proportion of examinees that answers a subset of questions correctly
   c) the proportion of examinees that achieves a given cutoff score
   d) the proportion of questions answered correctly by an examinee
   e) the mean score achieved by an examinee
   f) the percentile rank of a mean on a national norm distribution
   g) the percentile rank of a group percentile, referenced to a national norm distribution
3) The universe of measurement content:
   a) a single question on a measurement instrument
   b) all questions on the measurement instrument administered
   c) questions on the measurement instrument administered that are used

to assess mastery of an objective

d) all questions that could be used to assess mastery of an objective

e) all questions that could be used to assess status in a content domain

4) The universe of examinee generalization--all examinees of a given age or grade in the administrative unit(s) designated:

a) the state

b) each school system in the state

c) each school in the state

d) each classroom in the state

e) each school in a school system

f) each classroom in a school system

5) The procedure used to select examinees for assessment:

a) measurement of all examinees in the universe of interest

b) measurement of a simple random sample of examinees

c) measurement of a stratified sample of examinees

d) measurement of all examinees in a simple random sample of classrooms

e) measurement of all examinees in a simple random sample of schools

f) measurement of all examinees in a simple random sample of school systems

Neither the dimensions nor the categories of the taxonomy provided above are claimed to be exhaustive. They represent combinations of factors that describe assessments conducted in several states during the past four years (e.g., Pennsylvania, Florida, Oregon and California) and assessment procedures judged to be of potential interest.

A given universe score and a corresponding observed score are completely described by selecting a category from each taxonomic dimension. However, some combinations of categories provide universe scores unlikely to be of interest, and other combinations may be logically inconsistent. An example of a logically consistent universe score-observed score combination is provided by categories 1-b), 2-a), 3-a), 4-b), 5-b). Here the universe score would be the proportion of examinees in each school system in the state that can answer a particular question correctly. The interpretation of this universe score would be referenced to the domain of content from which the question was selected. The observed score would be a sample proportion for each school system, based on measurement of a simple random sample of examinees in each school system in the state. Considering only combinations of examinee universes and examinee selection procedures that are logically consistent and feasible, the taxonomy generates 595 different situations.

Generalized Indices of Stability. The most widely used indices of the reliability of individual scores follow Spearman's (1904) definition: "the average correlation between one and another of...several independently obtained series of values for p." However, as was illustrated above, correlations of successive observed scores for institutions may not provide stability indices that can be interpreted in useful or unambiguous ways. Three alternative indices of stability are suggested here.

In the literature on sampling from finite populations, the most widely used indicator of the stability of a statistic is its standard error; that is, the standard deviation of the sampling distribution of the statistic. If the sampling distribution of a statistic is known (or better yet, if the central limit theorem can be invoked), confidence statements can be constructed using the value of the statistic and its standard error. If ϕ is a universe score of interest, and q is an estimator of ϕ with a distribution that is asymptotically normal, an approximate $100(1-\alpha)$ percent confidence interval on ϕ is of the form $q \pm {}_{1-\alpha/2}z\sigma_q$

where

$_{1-\alpha/2}z$ denotes the $100(1-\alpha/2)$ percentile of the standard normal distribution, and σ_q denotes the standard error of q.

The standard error of the observed score used as an estimator of the universe socre is thus suggested as an indicator of stability for use in statewide assessments.

An advantage of the standard error is that its magnitude is expressed in the same units as those of the observed score it describes. Thus if the observed score is the mean raw score on an achievement test for a random sample of third-graders in a school system, the standard error of the mean will also be expressed in raw-score points. For some purposes, this otherwise convenient feature of the standard error can be troublesome. For example, if the stabilities of two measurement procedures that used different instruments were to be compared, direct comparison of respective standard errors would not, in general, be appropriate. In most instances, one unit on the scale of measurement of one instruemnt would not equal one unit on the scale of measurement of another instrument. A useful feature of Spearman's reliability index is its lack of dependence on the units of the measurement instrument it describes.

An alternative indicator of the stability of a statistic that has the "unitless" property of the Spearman reliability coefficient is the coefficient of variation (cv). The coefficient of variation is a discriptor sometimes used in the theory of sampling from finite populations. It is equal to the ratio of the standard error of a statistic to the value of the statistic. Thus for an observed score q with standard error σ_q, the coefficient of variation equals

$$cv(q) = \sigma_q/q.$$

The coefficients of variation of the observed scores on two different measurement instruments can be directly compared, without reference to the units of measurement of either instrument. The larger the coefficient of variation, the less stable the estimate of the universe score of interest.

To be consistent with the traditions of reliability estimation, it may be desirable to have an index of stability, rather than an indicator of the instability of an observed score. Such an index can be readily constructed from the coefficient of variation as follows:

Define the index of stability (IS) of an observed score q, used as an estimator of a universe score ϕ, to be

$$IS(q) = [1 - cv(q)]100 = [1 - \sigma_q/q]100 \text{ percent.}$$

Using this definition, the index of stability of an observed score equals 100 percent only if the standard error of the score, across all elements of the universe, equals zero. The index of stability equals zero if the standard error of the observed score is equal in magnitude to the observed score (Note that each of the observed-score statistics listed in dimension 2 of the taxonomy given above can only assume nonnegative values.) The index of stability assumes negative values only when the standard error of an observed score is larger than the value of the observed score.

Universe Scores, Observed Scores and their Estimated Standard Errors. Each combination of factors in the taxonomy provided above leads to a universe score, a corresponding observed score, and a standard error of the observed score. In order to estimate the stability of an observed score using the indices suggested in the preceeding section, an estimate of the standard error of each type of observed score must be available.

The combinations of universes of examinee generalization, examinee sampling procedures, observed scores, and universes of measurement generalization that are logically consistent and likely to be of some interest in a statewide assessment provide 595 entries in the previously described taxonomy. If the suggested indices of stability are to be computed, an estimator (formula for computing an observed score) and an estimate of standard error is needed for each of these entries. Although estimators and standard errors are nearly identical for some entries in the taxonomy, examination of all 595 cases is beyond the scope of this paper. Only the 170 cases generated by observed score entry 2a) "The proportion of examinees that answers a question correctly" and entry 2b) "The proportion of examinees that answers a subset of questions correctly" have been investigated.

An index to estimators of universe scores and corresponding standard errors is provided in Table 1. Entries in this table define components of estimators, and reference specific formulas provided in Table 2. As an example, suppose the evaluative referent for interpretation of the universe score of interest is a domain of content or abilities (Category 1a), the type of statistic that constituted the observed score is the proportion of examinees that answers a subset of questions correctly (Category 2b), the universe of measurement content is all questions on the measurement instrument administered (Category 3b), the universe of examinee generalization is composed of all examinees of a given age or grade in the state (Category 4a), and the procedure used to select examinees is simple random sampling (Category 5b). Reference to the appropriate estimator of the universe score of interest, and the standard error of the estimator, can then be found in Table 1 as Equation (10). Additional information in Table 1, needed to use Equation (10) for the specified purpose, includes the following definitions: N denotes the size of the examinee population in the state, n denotes the size of the examinee sample, M denotes the number of questions on the measurement instrument to which generalization is desired, and m denotes the number of questions sampled from the measurement instrument.

After determining the appropriate universe-score estimator and standard error for a given purpose, the user would turn to Table 2 to find the needed formulas. The parameters of formulas given in Table 2 would be defined using the information given in the appropriate cell of Table 1. In the example under discussion, the estimator of the universe score of interest is the sample proportion of examinee-question contacts that result in a correct answer ($p=f/nm$), where f is the number of examinee-question contacts that result in a correct answer. The standard error of p is given by the expression that follows S(p).

No derivations of estimators or standard errors are provided in this paper. Some of the results provided are new (particularly for stratified sampling of examinees), but most have been adapted and extended from the writings of Jaeger (1970), Lord and Novick (1968, Chapter 11), and Sirotnik (1974). Detailed derivation of many results involving matrix sampling may be found in Lord and Novick, and Sirotnik derives estimators of standard errors for several matrix sampling problems. Derivations of many results involving one-dimensional sampling (sampling either examinees or questions, but not both) follow developments provided in Cochran (1963).

In all cases where generalizations from a subset of questions to a subsuming set of questions is to be made, it is assumed that elements of the subset are selected through simple random sampling. When generalizations are to be made to "all questions that could be written," it is assumed that the universe of questions is infinite in size. In situations involving stratified sampling of examinees, it is assumed that examinees are selected using independent simple random sampling procedures within each stratum. Finally, in situations involving cluster sampling (e.g., sampling of classrooms, schools or school systems), it is assumed that clusters are selected using simple random sampling, that clusters may be unequal in size, and that all examinees within a sampled cluster are measured (single-stage cluster sampling). The estimators suggested for use in cluster sampling situations provide unbiased estimation of the universe scores of interest, and other than cluster sizes, require no information on the populations within clusters. Alternative single-stage cluster sampling and estimation procedures may provide more efficient estimation of universe scores, but these procedures require auxiliary information on examinees in the population (see Jaeger, 1970, 1973 for details).

For a number of the situations identified in Table 1, no analytic solution for the standard error of the estimated universe score is known. For these cases, the reader is referred to Table 2 for an unbiased estimator of the universe score of interest, and is instructed to use the jackknife procedure to estimate the standard error. Details on the application of the jackknife procedure can be found in Miller (1964), Mosteller and Turkey (1968, and Jaeger (1970).

Indicators of Cultural Validity

An achievement measure can be termed "culturally valid" provided groups or individuals of equal ability have the same chance of performing successfully. Cole (1973) proposed a philosophically similar definition of tests free from bias, when used for purposes of selection. Alternative definitions of culturally fair tests have been provided by Thorndike (1971) Darlington (1971), Einhorn and Bass (1971), and Linn (1973).

Since the measures used in statewide assessments are administered to individuals from diverse cultural and racial groups, and to groups of diverse cultural and racial composition, cultural validity (in the sense of being free from cultural or racial bias) is important.

To apply the definition of cultural validity proposed above, one must have a culturally-fair measure of ability to use as a standard. In statewide assessments (as in many other measurement situations), such culturally-fair ability measures are unlikely to exist. Thus it may be impossible to determine the cultural validity of an entire measurement instrument in these applications. However, it is feasible to determine the degree to which components of a measurement instrument provide relative cultural validity, provided performance on the entire instrument is accepted as a standard of ability. Cardall and Coffman (1964) and Cleary and Hilton (1968) examined the relative cultural bias of items on particular measurement instruments (the Scholastic Aptitude Test and the Preliminary Scholastic Aptitude Test, respectively) using Type III analysis of variance models. They used the item-by-group interaction component of variance as an indicator of the degree of relative cultural bias in the collective items of an instrument. The Cardall and Cleary procedures did not permit the identification of specific items that contributed to relative cultural bias; they only provided a global indicator of relative bias for the entire instrument.

Another procedure for estimating the relative cultural validity in the components of a measurement instrument was proposed by Angoff and Sharon (1974). Their method provided indicators of relative cultural validity for each item, in addition to providing an overall index for the instrument. Angoff and Sharon computed a normalized transformation of item difficulties for members of each of two cultural groups, used these to construct a scatter plot, and computed the equation of the major axis of the elliptical envelope surrounding the scatter of points. For each item, the perpendicular distance from the major axis was used as an indicator of relative cultural bias. Both the analysis of variance procedure and the Angoff and Sharon procedure might be said to employ the definition of cultural validity proposed here, but each uses the score on an entire measurement instrument as a measure of ability.

Although the Angoff and Sharon procedure was proposed for use with a norm-referenced achievement test, it could be applied to items that compose a domain-referenced or objectives-referenced measure. In fact, the procedure should provide interpretable results when used with any set of items that are homogeneous in content or purpose.

An alternative indicator of the relative cultural validity of items that compose a measure can be developed using empirical item characteristic functions (Henrysson, 1971). Again, the indicator is philosophically consistent with Cole's conditional probability model (1973) of culture fairness. Examinees' scores on the entire achievement measure are used as a surrogate for a culture-fair ability measure. The resulting sacrifice is an absolute index of cultural validity; only an index of relative validity can be obtained.

If the relative cultural validity of an item for members of two groups is to be estimated (call the groups Group A and Group B), proceed as follows. For each group, compute the cumulative distribution of the proportion of examinees who are successful on the item of interest, as a function of total score on items in a subsuming content domain (or on a subsuming test). For example, suppose a domain-referenced measure of ability to recognize sound blends contains 10 items, and the relative cultural validity of an item requiring identification of the "bl" sound blend is to be determined. The cumulative proportion of examinees from Group A who successfully identify "bl" and who earn scores of zero, one or less, two or less, . . ., on the entire sound blend item sample would be computed. (this is an empirical item characteristic function). The same procedure would be completed for examinees in B. As a measure of the relative cultural validity of the "bl" item, the maximum difference between the two empirical item characteristic functions would be computed. The smaller the maximum difference, the greater the relative cultural validity of the item. Fictitious item characteristic functions are illustrated in Figure 1, below.

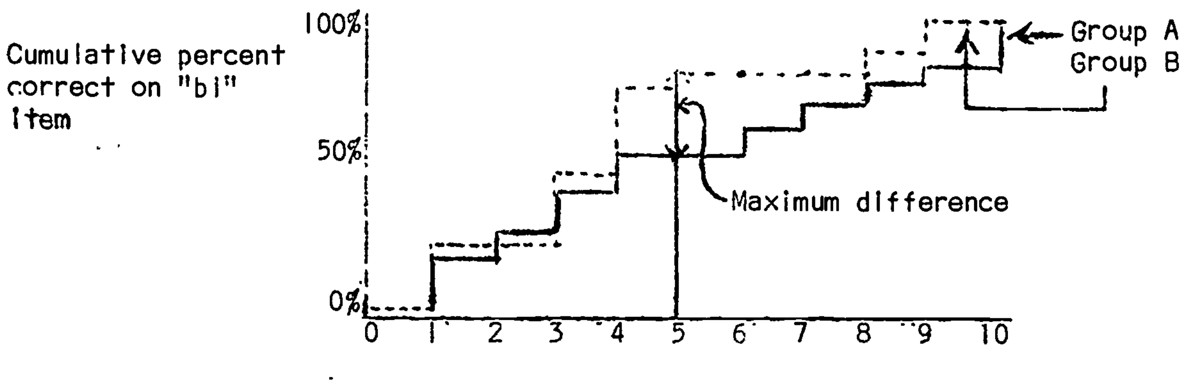

Figure 1: Fictitious Empirical Item Characteristic Functions

In the example portrayed by the graphs in Figure 1, the maximum difference between the item characteristic functions occurs at a score value of 5, and equals about 30 percent. In this example then, the "bl" item would have very low relative cultural validity, being relatively biased against members of Group A. Inspection of the curves would show that members of Group A who correctly answer about 50 percent of the items sampled from the consonant blends domain suffer the greatest relative bias on the "bl" item.

The procedure for estimating relative cultural validity proposed here would appear to identify those items that are relatively invalid, as does the Angoff and Sharon (1974) procedure, and in addition, identify the overall performance level of group members who suffer the greatest relative bias. This additional information may be useful when one seeks to determine why some items show relative cultural bias.

Procedures for Investigating Criterion Validity

In some statewide assessments, a pupil is said to have mastered an objective if (s)he successfully answers k out of K items related to that objective. The items related to a specific objective are said to constitute an exercise.

The problem of establishing a valid mastery score (k) is herein called the "criterion validity" problem.

In a recent review of research, Millman (1973) describes five procedures for establishing mastery scores on objectives-referenced exercises. Several of the procedures appear to be so administratively cumbersome as to be impractical, and all require subjective judgment at some point in their application. In short, there appears to be no universally best solution to this problem.

The procedure proposed here also depends upon subjective judgments, but identifies exercises for which independent judgments are inconsistent. Its strength then, lies in its consistency requirements.

For random samples of several hundred pupils in each grade assessed, teachers could be asked to specify whether or not each pupil has achieved minimal mastery in each content domain assessed. Data should be collected in such a way that teachers' judgments of minimal mastery can be matched to pupils' performance on assessment exercises.

In addition to providing these data, a small sample of randomly selected teachers should be asked to subjectively judge the difficulty of exercises associated with each objective in the domains assessed in their grades. These teachers would be asked to provide two types of judgments. First, they would be asked to estimate the proportions of pupils who have and have not achieved mastery of the content domain who should successfully answer 1 out of K, 2 out of K, . . ., K out of K items used to assess mastery of an objective. Second, these teachers would be asked to relate the difficulty of the exercises, by estimating the proportion of pupils who should be able to master a given objective, given that they have demonstrated mastery of another objective.

Using these data, the consistency of actual performance and estimated performance on each exercise could be examined. The proportions of pupils identified as masters and non-masters (a) that can successfully answer 1 out of K, 2 out of K, . . ., K out of K items in an exercise and (b) that teachers judge should be able to successfully answer 1 out of K, . . ., K out of K items in an exercise, could be compared. Large discrepancies between actual and judged percentages will pinpoint exercises for which prescribed mastery levels are inconsistent with independent judgments.

Teachers judgements of the conditional proportions of pupils who should exhibit mastery of one objective, given mastery of another, can be compared to actual proportions. Again, inconsistencies will pinpoint exercises for which mastery levels should be reconsidered.

Finally, considering once again the prospect that exercises can be classified into logical content domains, one could examine criterion validity by analyzing responses to all exercises in a content domain. If there are N_E exercises in a content domain, the proportion of pupils who have achieved mastery of a given exercise in the domain, and who have achieved mastery of 1 out of N_E exercises, 2 out of N_E exercises, . . ., N_E out of N_E exercises in the domain, can be tabulated. Any deviation from a monotonically increasing function would indicate inconsistency in prescribed mastery levels.

Summary

The most frequent interpretations of data from statewide assessments are institutional, rather than individual. To judge the quality of such interpretations, indices of stability and validity appropriate for institutional assessments are needed. Just as traditional psychometric indices prescribe upper bounds on the quality of interpretations of individual measurements, appropriate for institutional measurements will provide much needed warning flags.

This paper provides several indices of psychometric quality for institutional interpretations of measurement likely to be found in statewide assessments. However, the situations for which indices of stability are provided, and the limited types of measurement validity considered make but a small dent in a problem of major scope. Standard errors of observed scores must be developed for the 425 cells of the taxonomy not considered in this paper. And certainly, the taxonomy presented does not exhaust the situations that may arise in statewide assessments or other institutional uses of measurement. In the area of validity, progress has barely begun. As the purposes of statewide assessment are more clearly delineated, needs for validity indices parallel to those used with individual problems--the validity of domain representations, predictive validity, concurrent validity--will come into focus.

REFERENCES

Angoff, W. H. and A. T. Sharon "The evaluation of differences in test performance of two or more groups:, Educational and Psychological Measurement, 34, (1974), pp. 807-816.

Cardall, C. and W. E. Coffman "A method for comparing the performance of different groups on the items in a test", Research Bulletin No. 64-61, Princeton, New Jersey: Educational Testing Service, 1964.

Cleary, T. A. and T. L. Hilton "An investigation of item bias", Educational and Psychological Measurement, 28, (1968), pp. 61-75.

Cochran, W. G. Sampling Techniques, New York: John Wiley and Sons, 1963.

Cole, Nancy, "Bias in selection" Journal of Educational Measurement, 10, 4, (1973), pp. 237-256.

Connecticut State Legislature, Public Act Number 665, Hartford, Connecticut, 1971, 2 pp.

Cronback, L. J. "Coefficient alpha and the internal structure of tests:, Psychometrika, 16, (1951), pp. 297-334.

Cronbach, L. J., G. C. Gleser, H. Nanda, and N. Rajaratnum The Dependability of Behavioral Measurements, New York: John Wiley and Sons, 1972.

Darlington, R. B. "Another look at cultural fairness'", Jounral of Educational Measurement, 8, (1971), pp. 71-82.

Ebel, R.L. "Evaluation and educational objectives", Journal of Educational Measurement, 10, 4, (1973), pp. 278-89.

Einhorn, N. J. and A. R. Bass "Methodological considerations relevant to discrimination in employment testing", Psychological Bulletin, 75, (1971), pp. 261-269.

Georgia State Legislature, Senate Bill Number 672, Atlanta, Georgia, 1974, 70 pp.

Harris, C. "An interpretation of Livingston's reliability coefficient for criterion-referenced tests , Journal of Educational Measurement, 9, 1, (1972), pp. 27-30.

Harris, C. "Note on the variances and covariances of three error types", Journal of Educational Measurement, 10, 1, (1973), pp. 49-50.

Hawthorne, Phyllis Annotated Bibliography of the State Educational Accountability Repository, Denver, Colorado: Report Number 1, Cooperative Accountability Project

Henrysson, S. "Gathering, analyzing and using data on test items", in Thorndike, R. L., (ed) Educational Measurement 2nd ed., Washington, D.C.: American Council on Education, (1971), pp. 146ff.

Jaeger, R. M. "Designing school testing programs for institutional appraisal: an application of sampling theory", Unpublished doctoral dissertation, Stanford University, 1970.

Jaeger, R. M. "An evaluation of sampling designs for school testing programs", presented at the Annual Meeting of the National Council on Measurement in Education, 1973, New Orleans, LA.

Linn, R. L. "Fair test use in selection", Review of Educational Research, 43, (1973), pp. 139-161.

Livingston, S. W. "Criterion-referenced applications of classical test theory", Journal of Educational Measurement, 9, 1, (1972a), pp. 13-26.

Livingston, S. W. "A reply to Harris's 'An interpretation of Livingston's reliability coefficient'", Journal of Educational Measurement, 9, 1, (1972b), pp. 31-32.

Lord, F. M., and M. R. Novick Statistical Theories of Mental Test Scores, Reading, MA: Addison-Wesley Publishing Co., 1968.

Miller, R. G. "A trustworthy jackknife", Annals of Mathematical Statistics, 35, (1964), pp. 1594-1605.

Millman, J. "Passing scores and test lengths for domain-referenced measures", Review of Educational Research, 43, 2, (1973), pp. 205-216.

Mosteller, F., and J. W. Tukey "Data analysis, including statistics", in G. Lindzey and E. Aronson (Eds.), The Handbook of Social Psychology, Vol. II, Reading, MA: Addison-Wesley, (1968), pp. 80-203.

Pennsylvania State Legislature, School District Reorganization Act of 1963, Section 290.1, Harrisburg, PA, 1 p.

Popham, J. W, and T. R. Husek "Implications for criterion-referenced measurement", Journal of Educational Measurement, 6, (1969), pp. 1-9.

Sirotnik, K. A. "Introduction to matrix sampling for the practitioner", in W. J. Popham, (Ed.), Evaluation in Education, Current Applications, Berkeley, California: McCutchan Publishing Co., (1974), pp. 453-529.

Spearman, Charles "The proof and measurement of association between two things", American Journal of Psychology, 15, (1904), pp. 72-101.

Stanley, J. C. "Reliability", in R. L. Thorndike (Ed.), Educational Measurement, 2nd. ed., Washington, D.C.: American Council on Education, 1971.

Thorndike, R.L. "Concepts of culture-fairness", Journal of Educational Measurement, 8, (1971), pp. 63-70.

U.S. Government Printing Office. Equivalence and Norms Tables for Selected Reading Achievement Tests, Washington, D.C., 1974.

Table 1

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES*

<table>
<tr><td colspan="2">1) Evaluative Referent</td><td colspan="2">a) Domain of content or abilities</td></tr>
<tr><td colspan="2">2) Type of Statistic - Observed Score</td><td colspan="2">a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)</td></tr>
<tr><td colspan="2">3) Universe of Measurement Content</td><td rowspan="2">a) A single question on a measurement instrument</td><td rowspan="2">b) All questions on the measurement instrument administered</td></tr>
<tr><td>4) Universe of Examinee Generalization</td><td>5) Selection Procedure</td></tr>
<tr><td rowspan="3">a) The State</td><td>a) All examinees in state</td><td>Standard Error equals zero</td><td>Standard error is not estimable unless questions are of equal difficulty; then use Equation (6).
N=pop. size in state
M=No. questions on meas. instrument</td></tr>
<tr><td>b) SRS examinees in state</td><td>Equation (1)
N=pop. size in state
n=sample size in state
p=sample proportion that answer correctly</td><td>Standard error is not estimable unless questions are of equal difficulty; then use Equation (8).
N=pop. size in state
n=examinee sample size
M=No. questions on meas. instrument</td></tr>
<tr><td>c) Stratified sample of examinees in state</td><td>Equation (2)
N=pop. size in state
K=No. of strata
N_k= pop. size in stratum k
n=sample size
p_k=sample proportion in stratum k</td><td>Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases.</td></tr>
</table>

*See Table 2 for numbered estimators of standard errors.

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	
3) Universe of Measurement Content		a) A Single question on a measurement instrument	b) All questions on the measurement instrument administered
4) Universe of Examinee Generalization	5) Selection Procedure		
a) The State	d) SRS classrooms in state	Equation (3) N=No. classrooms in state n=No. classrooms in sample M_o=pop. examinees in state classroom=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. classrooms in state n=No. classrooms in sample M_o=pop. examinees in state classroom=cluster
	e) SRS schools in state	Equation (3) N=No. schools in state n=No. schools in sample M_o=pop. size examinees in state school=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. schools in state n=No. schools in sample M_o=pop. size examinees in state school=cluster

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	
3) Universe of Measurement Content		a) A single question on a measurement instrument	b) All questions on the measurement instrument administered
4) Universe of Examinee Generalization	5) Selection Procedure		
a) The State	f) SRS school systems in state	Equation (3) N=No. systems in state n=No. systems in sample M_o=pop. size examinees in state school system=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. systems in state n=No. systems in sample M_o=pop. size examinees in state school system=cluster
b) Each school system in state	a) All examinees in school system	Standard Error equals zero	Standard error is not estimable unless questions are of equal difficulty; then use Equation (6). N=pop. size in system M=No. questions on meas. instrument
	b) SRS examinees in school system	Equation (1) N=pop. size in system n=sample size in system p=sample proportion that answer correctly	Standard error is not estimable unless questions are of equal difficulty; then use Equation (8). N=pop. size in system n=examinee sample size M=No. questions on meas. instrument

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	
3) Universe of Measurement Content		a) A single question on a measurement instrument	b) All questions on the measurement instrument administered
4) Universe of Examinee Generalization	5) Selection Procedure		
b) Each school system in state	c) Stratified sample of examinees in school system	Equation (2) N=pop. size in system K=No. strata N_k=pop. size in stratum k n=sample size p_k=sample proportion in stratum k	Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases.
	d) SRS classrooms in school system	Equation (3) N=No. classrooms in system n=No. classrooms in sample M_o=pop. size examinees in system classroom=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\overline{p}$ and jackknife procedure to estimate standard error. N=No. classrooms in system n=No. classrooms in sample M_o= pop. size examinees in system classroom=cluster

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	
3) Universe of Measurement Content		a) A single question on a measurement instrument	b) All questions on the measurement instrument administered
4) Universe of Examinee Generalization	5) Selection Procedure		
b) Each school system in state	e) SRS schools in system	Equation (3) N=No. schools in system n=No. schools in sample M_o=pop. size examinees in system school=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. schools in system n=No. schools in sample M_o=pop. size examinees in system school=cluster
c) Each school in state or e) Each school in a system	a) All examinees in school	Standard Error equals zero	Standard error is not estimable unless questions are of equal difficulty; then use Equation (6). N=pop. size in school M=No. questions on meas. instrument
	b) SRS examinees in school	Equation (1) N=pop. size in school n=sample size in school p=sample proportion in school that answer correctly	Standard error is not estimable unless questions are of equal difficulty; then use Equation (8). N=pop. size in school n=examinee sample size M=No. questions on meas. instrument

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	
3) Universe of Measurement Content		a) A single question on a measurement instrument	b) All questions on the measurement instrument administered
4) Universe of Examinee Generalization	5) Selection Procedure		
c) Each school in state or e) Each school in a system	c) Stratified sample of examinees in school	Equation (2) N=pop. size in school K=No. of strata N_k=pop. size in stratum k n=sample size p_K= sample proportion in stratum k	Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases.
	d) SRS classrooms in school	Equation (3) N=No. classrooms in school n=No. classrooms in sample M_o=pop. examinees in school classroom=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solution is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. classrooms in school n=No. classrooms in sample M_o=pop. examinees in school classroom=cluster
d) Each classroom in state or f) Each classroom in a system	a) All examinees in classroom	Standard Error equals zero	Standard error is not estimable unless questions are of equal difficulty; then use Equation (6). N=pop. size in class-[illegible] M=No. questions on meas. instrument

Table 1

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

<table>
<tr><td colspan="2">1) Evaluative Referent</td><td colspan="2">a) Domain of content or abilities</td></tr>
<tr><td colspan="2">2) Type of Statistic - Observed Score</td><td colspan="2">a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)</td></tr>
<tr><td colspan="2">3) Universe of Measurement Content</td><td rowspan="2">a) A single question on a measurement instrument</td><td rowspan="2">b) All questions on the measurement instrument administered</td></tr>
<tr><td>4) Universe of Examinee Generalization</td><td>5) Selection Procedure</td></tr>
<tr><td>d) Each classroom in state
or
f) Each classroom in a system</td><td>b) SRS examinees in classroom</td><td>Equation (1)
N=pop. size in classroom
n=sample size in classroom
p=sample proportion that answer correctly</td><td>Standard error is not estimable unless questions are of equal difficulty; then use Equation (8).
N=pop. size in classroom
n=examinee sample size
M=No. questions on meas. instrument</td></tr>
</table>

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	
3) Universe of Measurement Content		c) Questions used to assess mastery of an objective	d) Questions that could be used to assess mastery of an objective
4) Universe of Examinee Generalization	5) Selection Procedure		
a) The State	a) All examinees in state	Standard error is not estimable unless questions are of equal difficulty; then use Equation (6). N=pop. size in state M=No. questions that pertain to objective	Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (7). N= pop. size in state
	b) SRS examinees in state	Standard error is not estimable unless questions are of equal difficulty; then use Equation (8). N=pop. size in state n=examinee sample size M=No. questions that pertain to objective	Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (9). n=examinee sample size
	c) Stratified sample of examinees in state	Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases.	Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases.

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	
3) Universe of Measurement Content		c) Questions used to assess mastery of an objective	d) Questions that could be used to assess mastery of an objective
4) Universe of Examinee Generalization	5) Selection Procedure		
a) The State	d) SRS classrooms in state	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\overline{p}$ and jackknife procedure to estimate standard error. N=No. classrooms in state n=No. classrooms in sample M_o=pop. examinees in state classroom=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\overline{p}$ and jackknife procedure to estimate standard error. N=No. classrooms in state n=No. classrooms in sample M_o=pop. examinees in state classroom=cluster
	e) SRS schools in state	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\overline{p}$ and jackknife procedure to estimate standard error. N=No. schools in state n=No. schools in sample M_o=pop. size examinees in state school=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\overline{p}$ and jackknife procedure to estimate standard error. N=No. schools in state n=No. schools in sample M_o=pop. size examinees in state school=cluster

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	
3) Universe of Measurement Content		c) Questions used to assess mastery of an objective	d) Questions that could be used to assess mastery of an objective
4) Universe of Examinee Generalization	5) Selection Procedure		
a) The State	f) SRS school systems in state	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\overline{p}$ and jack-knife procedure to estimate standard error. N=No. systems in state n=No. systems in sample M_o=pop. size examinees in state school system=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\overline{p}$ and jack-knife procedure to estimate standard error. N=No. systems in state n=No. systems in sample M_o=pop. size examinees in state school system=cluster
b) Each school system in state	a) All examinees in school system	Standard error is not estimable unless questions are of equal difficulty; then use Equation (6). N=pop. size in system M=No. questions that pertain to objective	Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (7). N=pop. size in system

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

<table>
<tr><td colspan="2">1) Evaluative Referent</td><td colspan="2">a) Domain of content or abilities</td></tr>
<tr><td colspan="2">2) Type of Statistic - Observed Score</td><td colspan="2">a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)</td></tr>
<tr><td colspan="2">3) Universe of Measurement Content</td><td rowspan="2">c) Questions used to assess mastery of an objective</td><td rowspan="2">d) Questions that could be used to assess mastery of an objective</td></tr>
<tr><td>4) Universe of Examinee Generalization</td><td>5) Selection Procedure</td></tr>
<tr><td rowspan="2">b) Each school system in state</td><td>b) SRS examinees in school system</td><td>Standard error is not estimable unless questions are of equal difficulty; then use Equation (8).
N=pop. size in system
n=examinee sample size
M=No. questions that pertain to objective</td><td>Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (9).
n=examinee sample size</td></tr>
<tr><td>c) Stratified sample of examinees in school system</td><td>Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases.</td><td>Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases.</td></tr>
</table>

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic – Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	
3) Universe of Measurement Content		c) Questions used to assess mastery of an objective	d) Questions that could be used to assess mastery of an objective
4) Universe of Examinee Generalization	5) Selection Procedure		
b) Each school system in state	d) SRS classrooms in school system	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure t estimate standard error. N=No. classrooms in system n=No. classrooms in sample M_o=pop. size examinees in system classroom=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solutions for standard error is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. classrooms in system n=No. classrooms in sample M_o=pop. size examinees in system classroom=cluster
	e) SRS schools in system	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. schools in system n=No. schools in sample M_o=pop. size examinees in system school=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. schools in system n=No. schools in sample M_o=pop. size examinees in system school=cluster

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	
3) Universe of Measurement Content		c) Questions used to assess mastery of an objective	d) Questions that could be used to assess mastery of an objective
4) Universe of Examinee Generalization	5) Selection Procedure		
c) Each school in state or e) Each school in a system	a) All examinees in school	Standard error is not estimable unless questions are of equal difficulty; then use Equation (6). N=pop. size in school M=No. questions that pertain to objective	Standard error is not estimable unless questions are of equal difficulty; then use Equation (7). N=pop. size in school
	b) SRS examinees in school	Standard error is not estimable unless questions are of equal difficulty; then use Equation (8). N=pop. size in school n=examinee sample size M=No. questions that pertain to objective	Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (9). n=examinee sample size
	c) Stratified sample of examinees in school	Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases.	Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases.

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	
3) Universe of Measurement Content		c) Questions used to assess mastery of an objective	d) Questions that could be used to assess mastery of an objective
4) Universe of Examinee Generalization	5) Selection Procedure		
c) Each school in state or e) Each school in a system	d) SRS classrooms in school	Standard error is not estimable unless questions are of equal difficulty; then analytic solution is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. classrooms in school n=No. classrooms in sample M_o=pop. examinees in school classroom=cluster	Standard error is not estimable unless questions are of equal difficulty; then analytic solution is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. classrooms in school n=No. classrooms in sample M_o=pop. examinees in school classroom=cluster
d) Each classroom in state or f) Each classroom in a system	a) All examinees in classroom	Standard error is not estimable unless questions are of equal difficulty; then use Equation (6). N=pop. size in classroom M=No. questions that pertain to objective	Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (7). N=pop. size in classroom
	b) SRS examinees in classroom	Standard error is not estimable unless questions are of equal difficulty; then use Equation (8). N=pop. size in classroom n=examinee sample size M=No. questions that pertain to objective	Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (9). n=examinee sample size

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

<table>
<tr><td colspan="2">1) Evaluative Referent</td><td colspan="2">a) Domain of content or abilities</td></tr>
<tr><td colspan="2">2) Type of Statistic - Observed Score</td><td>a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)</td><td>b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)</td></tr>
<tr><td colspan="2">3) Universe of Measurement Content</td><td rowspan="2">e) Questions that could be used to assess status in a content domain</td><td rowspan="2">a) A single question on a measurement instrument</td></tr>
<tr><td>4) Universe of Examinee Generalization</td><td>5) Selection Procedure</td></tr>
<tr><td rowspan="3">a) The State</td><td>a) All examinees in state</td><td>Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (7).
N=pop. size in state</td><td>Standard error is not estimable unless specific question is sampled; then standard error equals zero.</td></tr>
<tr><td>b) SRS examinees in state</td><td>Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (9).
n=examinee sample size</td><td>Standard error is not estimable unless specific question is sampled; then use Equation (1).
N=pop. size in state
n=sample size in state
p=sample proportion that answer correctly</td></tr>
<tr><td>c) Stratified sample of examinees in state</td><td>Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases</td><td>Standard error is not estimable unless specific question is sampled; then use Equation (2).
N=pop. size in state
K=No. of strata
N_k=pop. size in stratum k
p_k=sample proprotion in stratum k</td></tr>
</table>

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)
3) Universe of Measurement Content		e) Questions that could be used to assess status in a content domain	a) A single question on a measurement instrument
4) Universe of Examinee Generalization	5) Selection Procedure		
a) The State	d) SRS classrooms in state	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\overline{p}$ and jackknife procedure to estimate standard error. N=No. classrooms in state n=No. classrooms in sample M_o=pop. examinees in state classroom=cluster	Standard error is not estimable unless specific question is sampled; then use Equation (3). N=No. classrooms in state n=No. classrooms in sample M_o=pop. examinees in state classroom=cluster
	e) SRS schools in state	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\overline{p}$ and jackknife procedure to estimate standard error. N=No. schools in state n=No. schools in sample M_o=pop. size examinees in state school=cluster	Standard error is not estimable unless specific question is sampled; then use Equation (3). N=No. schools in state n=No. schools in sample M_o=pop. examinees in state school=cluster

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)
3) Universe of Measurement Content		e) Questions that could be used to assess status in a content domain	a) A single question on a measurement instrument
4) Universe of Examinee Generalization	5) Selection Procedure		
a) The State	f) SRS school systems in state	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. systems in state n=No. systems in sample M_o=pop. size examinees in state school system=cluster	Standard error is not estimable unless specific question is sampled; then use Equation (3). N=No. systems in state n=No. systems in sample M_o=pop. size examinees in state school system=cluster
b) Each school system in state	a) All examinees in school system	Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (7). N=pop. size in system	Standard error is not estimable unless specific question is sampled; then standard error equals zero.
	b) SRS examinees in school system	Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (9). n=examinee sample size	Standard error is not estimable unless specific question is sampled; then use Equation (1). N=pop. size in system n=sample size in system p=sample proportion that answer correctly

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)
3) Universe of Measurement Content		e) Questions that could be used to assess status in a content domain	a) A single question on a measurement instrument
4) Universe of Examinee Generalization	5) Selection Procedure		
b) Each school system in state	c) Stratified sample of examinees in school system	Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases.	Standard error is not estimable unless specific question is sampled; then use Equation (2). N=pop. size in school system K=No. of strata N_k=pop. size in stratum k p_k=sample proportion in stratum k
	d) SRS classrooms in school system	Standard error is not estimable unless questions are of equal difficulty; then analytic solution for standard error is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. classrooms in system n=No. classrooms in sample M_o=pop. size examinees in system classroom=cluster	Standard error is not estimable unless specific question is sampled; then use Equation (3). N=No. classrooms in system n=No. classrooms in sample M_o=pop. size examinees in system classroom=cluster

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)
3) Universe of Measurement Content		e) Questions that could be used to assess status in a content domain	a) A single question on a measurement instrument
4) Universe of Examinee Generalization	5) Selection Procedure		
b) Each school system in state	e) SRS schools in system	Standard error is not estimable unless questions are of equal difficulty; than analytic solution for standard error is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. schools in system n=No. schools in sample M_o=pop. size examinees in system school=cluster	Standard error is not estimable unless specific question is sampled; then use Equation (3). N=No. schools in system n=No. schools in sample M_o=pop. size examinees in system school=cluster
c) Each school in state or e) Each school in a system	a) All examinees in school	Standard error is not estimable unless questions are of equal difficulty; then use Equation (7). N=pop. size in school	Standard error is not estimable unless specific question is sampled; then standard error equals zero.

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)
3) Universe of Measurement Content		e) Questions that could be used to assess status in a content domain	a) A single question on a measurement instrument
4) Universe of Examinee Generalization	5) Selection Procedure		
c) Each school in state or e) Each school in a system	b) SRS examinees in school	Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (9). n=examinee sample size	Standard error is not estimable unless specific question is sampled; then use Equation (1). N=pop. size in school n=sample size in school p=sample proportion that answer correctly
	c) Stratified sample of examinees in school	Standard error is not estimable unless all questions are of equal difficulty for examinees within a stratum. This assumption is probably untenable in most cases.	Standard error is not estimable unless specific question is sampled; then use Equation (2). N=pop. size in school K=No. of strata N_k=pop. size in stratum k p_k=sample proportion in stratum k

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		a) Proportion of examinees that answer a single question correctly (Assume question is randomly sampled from relevant universe)	b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)
3) Universe of Measurement Content		e) Questions that could be used to assess status in a content domain	a) A single question on a measurement instrument
4) Universe of Examinee Generalization	5) Selection Procedure		
c) Each school in state or e) Each school in a system	d) SRS classrooms in school	Standard error is not estimable unless questions are of equal difficulty; then analytic solution is unknown, so use Equation (3) to estimate $\bar{p}$ and jackknife procedure to estimate standard error. N=No. classrooms in school n=No. classrooms in sample M_o=pop. examinees in school classroom=cluster	Standard error is not estimable unless specific question is sampled; then use Equation (3). N=No. classrooms in school n=No. classrooms in sample M_o=pop. examinees in school classroom=cluster
d) Each classroom in state or f) Each classroom in a system	a) All examinees in classroom	Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (7). N=pop. size in classroom	Standard error is not estimable unless specific question is sampled; then standard error equals zero.
	b) SRS examinees in classroom	Standard error is not estimable unless all questions in population are of equal difficulty; then use Equation (9). n=examinee sample size	Standard error is not estimable unless specific question is sampled; then use Equation (1) N=pop. size in classroom n=sample size in classroom p=sample proportion that answer correctly

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic		b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)	
3) Universe of Measurement Content		b) All questions on the measurement instrument administered	c) Questions used to assess mastery of an objective
4) Universe of Examinée Generalization	5) Selection Procedure		
a) The State	a) All examinees in state	Equation (4) M=No. questions on meas. instrument m=No. questions sampled N=pop. size in state	Equation (4) M=No. questions on instrument that pertain to objective m=No. sampled questions that pertain to objective N=pop. size in state
	b) SRS examinees in state	Equation (10) N=pop. size in state n=examinee sample size M=No. questions on instrument m=No. questions sampled	Equation (10) N=pop. size in state n=examinee sample size M=No. questions on instrument that pertain to objective m=No. sampled questions that pertain to objective
	c) Stratified sample of examinees in state	Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (12). N_i=pop. size in stratum i K=No. of strata n_i=sample size in stratum i M=No. questions on instrument m=No. questions sampled for each stratum	Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (12). N_i=pop. size in stratum i K=No. of strata n_i=sample size in stratum i M=No. questions on instrument that pertain to objective m=No. sampled questions that pertain to objective for each stratum

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)	
3) Universe of Measurement Content		b) All questions on the measurement instrument administered	c) Questions used to assess mastery of an objective
4) Universe of Examinee Generalization	5) Selection Procedure		
a) The State	d) SRS classrooms in state	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. classrooms in state n=No. classrooms in sample M_o=pop. examinees in state classroom=cluster	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. classrooms in state n=No. classrooms in sample M_o=pop. examinees in state classroom=cluster
	e) SRS schools in state	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. schools in state n=No. schools in sample M_o=pop. examinees in state school=cluster	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procudure to estimate standard error. N=No. schools in state n=No. schools in sample M_o=pop. examinees in state school=cluster
	f) SRS school systems in state	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procudure to estimate standard error. N=No. systems in state n=No. systems in sample M_o=pop. examinees in state school system=cluster	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procudure to estimate standard error. N=No. systems in state n=No. systems in sample M_o=pop. examinees in state school system=cluster

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)	
3) Universe of Measurement Content		b) All questions on the measurement instrument administered	c) Questions used to assess mastery of an objective
4) Universe of Examinee Generalization	5) Selection Procedure		
b) Each school system in state	a) All examinees in school system	Equation (4) N=pop. size in school system M=No. questions on meas. instrument m=No. questions sampled	Equation (4) N=pop. size in school system M=No. questions on instrument that pertain to objective m=No. sampled questions that pertain to objective
	b) SRS examinees in school system	Equation (10) N=pop. size in school system n=examinee sample size M=No. questions on instrument m=No. questions sampled	Equation (10) N=pop. size in school system n=examinee sample size M=No. questions on instrument that pertain to objective m=No. sampled questions that pertain to objective
	c) Stratified sample of examinees in school system	Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (12). N_i=pop. size in stratum i K=No. of strata n_i=sample size in stratum i M=No. questions on instrument m=No. questions sampled for each stratum	Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (12). N_i=pop. size in stratum i K=No. of strata n_i=sample size in stratum i M=No. questions on instrument that pertain to objective m=No. sampled questions that pertain to objective for each stratum

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)	
3) Universe of Measurement Content		b) All questions on the measurement instrument administered	c) Questions used to assess mastery of an objective
4) Universe of Examinee Generalization	5) Selection Procedure		
b) Each school system in state	d) SRS classrooms in school system	Analytic solution for standard error is unknown. Use Equation (3) to estimate $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. classrooms in system n=No. classrooms in sample M_o=pop. examinees in system classroom=cluster	Analytic solution for standard error is unknown. Use Equation (3) to estimate $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. classrooms in system n=No. classrooms in sample M_o=pop. examinees in system classroom=cluster
	e) SRS schools in system	Analytic solution for standard error is unknown. Use Equation (3) to estimate $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. schools in system n=No. schools in sample M_o=pop. examinees in sample school=cluster	Analytic solution for standard error is unknown. Use Equation (3) to estimate $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. schools in system n=No. schools in sample M_o=pop. examinees in sample school=cluster
c) Each school in state or e) Each school in a system	a) All examinees in school	Equation (4) N=pop. size in school M=No. questions on meas. instrument m=No. questions sampled	Equation (4) M=No. questions on instrument that pertain to objective m=No. sampled questions that pertain to objective N=pop. size in school

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

<table>
<tr><td colspan="2">1) Evaluative Referent</td><td colspan="2">a) Domain of content or abilities</td></tr>
<tr><td colspan="2">2) Type of Statistic - Observed Score</td><td colspan="2">b) Proportion of examimees that answer a subset of questions correctly (Assume simple random sampling of questions)</td></tr>
<tr><td colspan="2">3) Universe of Measurement Content</td><td rowspan="2">b) All questions on the measurement instrument administered</td><td rowspan="2">c) Questions used to assess mastery of an objective</td></tr>
<tr><td>4) Universe of Examinee Generalization</td><td>5) Selection Procedure</td></tr>
<tr><td rowspan="2">c) Each school in state
or
e) Each school in a system</td><td>b) SRS examinees in school</td><td>Equation (10)
N=pop. size in school
n=examinee sample size
M=No. questions on instrument
m=No. questions sampled</td><td>Equation (10)
N=pop. size in school
n=examinee sample size
M=No. questions on instrument that pertain to objective
m=No. sampled questions that pertain to objective</td></tr>
<tr><td>c) Stratified sample of examinees in school</td><td>Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (12).
N_i=pop. size in stratum i
K=No. of strata
n_i=sample size in stratum i
M=No. questions on instrument
m=No. questions sampled for each stratum</td><td>Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (12).
N_i=pop. size in stratum i
K=No. of strata
n_i=sample size in stratum i
M=No. questions on instrument that pertain to objective
m=No. sampled questions that pertain to objective for each stratum</td></tr>
</table>

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)	
3) Universe of Measurement Content		b) All questions on the measurement instrument administered	c) Questions used to assess mastery of an objective
4) Universe of Examinee Generalization	5) Selection Procedure		
c) Each school in state or e) Each school in a system	d) SRS classrooms in school	Analytic solution for standard error is unknown. Use Equation (3) to estimate $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. classrooms in school n=No. classrooms in sample M_o=pop. examinees in school classroom=cluster	Analytic solution for standard error is unknown. Use Equation (3) to estimate $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. classrooms in school n=No. classrooms in sample M_o=pop. examinees in school classroom=cluster
d) Each classroom in state or f) Each classroom in a system	a) All examinees in classroom	Equation (4) N=pop. size in classroom M=No. questions on meas. instrument m=No. questions sampled	Equation (4) M=No. questions on instrument that pertain to objective m=No. sampled questions that pertain to objective N=pop. size in classroom
	b) SRS examinees in classroom	Equation (10) N=pop. size in classroom n=examinee sample size M=No. questions on instrument m=No. questions sampled	Equation (10) N=pop. size in classroom n=examinee sample size M=No. questions on instrument that pertain to objective m=No sampled questions that pertain to objective

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)	
3) Universe of Measurement Content		d) Questions that could be used to assoss mastery of an objective	e) Questions that could be used to assess status in a content domain
4) Universe of Examinee Generalization	5) Selection Procedure		
a) The State	a) All examinees in state	Equation (5) Question population assumed infinite m=No. sampled questions that pertain to objective N=pop. size in state	Equation (5) Question population assumed infinite m=No. sampled questions N=pop. size in state
	b) SRS examinees in state	Equation (11) N=pop. size in state n=examinee sample size m=No. sampled questions that pertain to objective Question population assumed infinite	Equation (11) N=pop. size in state n=examinee sample size m=No. sampled questions Question population assumed infinite
	c) Stratified sample of examinees in state	Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (13). N_i=pop. size in stratum i K=No. of strata n_i=sample size in stratum i m=No. questions sampled for each stratum that pertain to objective	Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (13). N_i=pop. size in stratum i K=No. of strata n_i=sample size in stratum i m=No. questions sampled for each stratum

Table 1

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)	
3) Universe of Measurement Content		d) Questions that could be used to assess mastery of an objective	e) Questions that could be used to assess status in a content domain
4) Universe of Examinee Generalization	5) Selection Procedure		
a) The State	d) SRS classrooms in state	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. classrooms in state n=No. classrooms in sample M_o=pop. examinees in state classroom=cluster	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. classrooms in state n=No. classrooms in sample M_o=pop. examinees in state classroom=cluster
	e) SRS schools in state	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. schools in state n=No. schools in sample M_o=pop. examinees in state school=cluster	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. schools in state n=No. schools in sample M_o=pop. examinees in state school=cluster
	f) SRS school systems in state	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. systems in state n=No. systems in sample M_o=pop. examinees in state school system=cluster	Analytic solution for standard error is unknown. Use Equation (3) to compute $\bar{p}$, then use jackknife procedure to estimate standard error. N=No. systems in state n=No. systems in sample M_o=pop. examinees in state school system=cluster

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)	
3) Universe of Measurement Content		d) Questions that could be used to assess mastery of an objective	e) Questions that could be used to assess status in a content domain
4) Universe of Examinee Generalization	5) Selection Procedure		
b) Each school system in state	a) All examinees in school system	Equation (5) Question population assumed infinite m=No. sampled questions that pertain to objective N=pop. size in school system	Equation (5) Question population assumed infinite m=No. sampled questions N=pop. size in school system
	b) SRS examinees in school system	Equation (11) Question population assumed infinite N=pop. size in school system n=examinee sample size m=No. sampled questions that pertain to objective	Equation (11) Question population assumed infinite N=pop. size in school system n=examinee sample size m=No. sampled questions
	c) Stratified sample of examinees in school system	Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (13). N_i=pop. size in stratum i K=No. of strata n_i=sample size in stratum i m=No. questions sampled for each stratum that pertain to objective	Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (13). N_i=pop. size in stratum i K=No. of strata n_i=sample size in stratum i m=No. questions sampled for each stratum

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)	
3) Universe of Measurement Content		d) Questions that could be used to assess mastery of an objective	e) Questions that could be used to assess status in a content domain
4) Universe of Examinee Generalization	5) Selection Procedure		
b) Each school system in state	d) SRS classrooms in school system	Analytic solution for standard error is unknown. Use Equation (3) to estimate $\overline{p}$, then use jackknife procedure to estimate standard error. N=No. classrooms in system n=No. classrooms in sample M_o=pop. examinees in system classroom=cluster	Analytic solution for standard error is unknown. Use Equation (3) to estimate $\overline{p}$, then use jackknife procedure to estimate standard error. N=No. classrooms in system n=No. classrooms in sample M_o=pop. examinees in system classroom=cluster
	e) SRS schools in system	Analytic solution for standard error is unknown. Use Equation (3) to estimate $\overline{p}$, then use jackknife procedure to estimate standard error. N=No. schools in system n=No. schools in sample M_o=pop. examinees in sample school=cluster	Analytic solution for standard error is unknown. Use Equation (3) to estimate $\overline{p}$, then use jackknife procedure to estimate standard error. N=No. schools in system n=No. schools in sample M_o=pop. examinees in sample school=cluster
c) Each school in state or e) Each school in a system	a) All examinees in school	Equation (5) Question population assumed infinite m=No. sampled questions that pertain to objective N=pop. size in school	Equation (5) Question population assumed infinite m=No. sampled questions N=pop. size in school

Table I

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

<table>
<tr><td colspan="2">1) Evaluative Referent</td><td colspan="2">a) Domain of content or abilities</td></tr>
<tr><td colspan="2">2) Type of Statistic - Observed Score</td><td colspan="2">b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)</td></tr>
<tr><td colspan="2">3) Universe of Measurement Content</td><td rowspan="2">d) Questions that could be used to assess mastery of an objective</td><td rowspan="2">e) Questions that could be used to assess status in a content domain</td></tr>
<tr><td>4) Universe of Examinee Generalization</td><td>5) Selection Procedure</td></tr>
<tr><td rowspan="3">c) Each school in state
or
e) Each school in a system</td><td>b) SRS examinees in school</td><td>Equation (11)
Question population assumed infinite
N=pop. size in school
n=examinee sample size
m=No. sampled questions that pertain to objective</td><td>Equation (11)
Question population assumed infinite
N=pop. size in school
n=examinee sample size
m=No. sampled questions</td></tr>
<tr><td>c) Stratified sample of examinees in school</td><td>Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (13).
N_i=pop. size in stratum i
K=No. of strata
n_i=sample size in stratum i
m=No. sampled questions for each stratum that pertain to objective</td><td>Analytic solution is unknown unless questions are sampled independently for each examinee stratum; then use Equation (13).
N_i=pop. size in stratum i
K=No. of strata
n_i=sample size in stratum i
m=No. questions sampled for each stratum</td></tr>
<tr><td>d) SRS classrooms in school</td><td>Analytic solution for standard error is unknown. Use Equation (3) to estimate $\bar{p}$, then use jackknife procedure to estimate standard error.
N=No. classrooms in school
n=No. classrooms in sample
M_o=pop. examinees in school
classroom=cluster</td><td>Analytic solution for standard error is unknown. Use Equation (3) to estimate $\bar{p}$, then use jackknife procedure to estimate standard error.
N=No. classrooms in school
n=No. classrooms in sample
M_o=pop. examinees in school
classroom=cluster</td></tr>
</table>

Table 1

DESIGNATIONS OF ESTIMATORS OF STANDARD ERRORS OF OBSERVED SCORES
USED AS ESTIMATORS OF UNIVERSE SCORES

1) Evaluative Referent		a) Domain of content or abilities	
2) Type of Statistic - Observed Score		b) Proportion of examinees that answer a subset of questions correctly (Assume simple random sampling of questions)	
3) Universe of Measurement Content		d) Questions that could be used to assess mastery of an objective	e) Questions that could be used to assess status in a content domain
4) Universe of Examinee Generalization	5) Selection Procedure		
d) Each classroom in state or f) Each classroom in a system	a) All examinees in classroom	Equation (5) Question population assumed infinite m=No. sampled questions that pertain to objective N=pop. size in classroom	Equation (5) Question population assumed infinite m=No. sampled questions N=pop. size in classroom
	b) SRS examinees in classroom	Equation (11) Question population assumed infinite N=pop. size in classroom n=examinee sample size m=No. sampled questions that pertain to objective	Equation (11) Question population assumed infinite N=pop. size in classroom n=examinee sample size m=No. sampled questions

Table 2

ESTIMATORS OF UNIVERSE SCORES AND THEIR STANDARD ERRORS

ESTIMATOR	STANDARD ERROR
(1) f = number sampled who answer correctly ($p = f/n$) = (estimator)	$S(p) = \left[\frac{(N-n)}{(n-1)N}\, p(1-p)\right]^{1/2}$
(2) f_k = number sampled in stratum k who answer correctly n_k = sample size in stratum k $p_k = \frac{f_k}{n_k}$ ($p = \sum_{k=1}^{K} (N_{k}/N) p_k$) = (estimator)	$S(p) = \left[\sum_{k=1}^{K} \frac{N_k}{N^2} \cdot \frac{(N_k - n_k)}{(n_k - 1)}\, p_k(1-p_k)\right]^{1/2}$
(3) M_i = number of examinees in i-th cluster p_i = proportion of successes in i-th cluster $\bar{T} = \frac{1}{n} \sum_{i=1}^{n} M_i p_i$ $f = n/N$ ($\bar{p} = \frac{N}{nM_o} \sum_{i=1}^{n} M_i p_i$) = (estimator)	$S(\bar{p}) = \left[\frac{N}{M_o}\, \frac{(1-f)}{n(n-1)} \sum_{i=1}^{n} (M_i p_i - \bar{T})^2\right]^{1/2}$
(4) f = number of sampled question-examinee contacts that result in correct answers Nm = number of sampled question-examinee contacts NM = population size of question-examinee contacts ($p = f/Nm$) = (estimator)	$S(p) = \left[\frac{NM-Nm}{(Nm-1)NM}\, p(1-p)\right]^{1/2}$

Table 2

ESTIMATORS OF UNIVERSE SCORES AND THEIR STANDARD ERRORS

ESTIMATOR	STANDARD ERROR
(5) f = number of sampled question-examinee contacts that result in correct answers Nm = number of sampled examinee-question contacts (p = f/Nm) = (estimator)	$S(p) = \left[\frac{p(1-p)}{Nm - 1}\right]^{1/2}$
(6) f = number of sampled question-examinee contacts that result in correct answer. (p = f/N) = (estimator)	$S(p) = \left[\frac{NM - N}{(N-1)NM}\, p(1-p)\right]^{1/2}$
(7) f = number of sampled question-examinee contacts that result in correct answer. (p = f/N) = (estimator)	$S(p) = \left[\frac{p(1-p)}{N-1}\right]^{1/2}$
(8) f = number of sampled question-examinee contacts that result in correct answer (p = f/n) = (estimator)	$S(p) = \left[\frac{NM - n}{(n-1)NM}\, p(1-p)\right]^{1/2}$
(9) f = number of sampled question-examinee contacts that result in correct answer (p = f/n) = (estimator)	$S(p) = \left[\frac{p(1-p)}{n-1}\right]^{1/2}$

Table 2

ESTIMATORS OF UNIVERSE SCORES AND THEIR STANDARD ERRORS

ESTIMATOR	STANDARD ERROR
(10) f = number of sampled question-examinee contacts that result in correct answer (p = f/nm) = (estimator)	$S(p) = \left\{ \frac{1}{mn(M-1)(N-1)} \left[(M-m)\, n(N-1)\hat{\sigma}^2_I + m(M-1)(N-n)\hat{\sigma}^2_E + (M-m)(N-n)\hat{\sigma}^2_{EI} \right] \right\}^{1/2}$

where variance component estimators are defined as follows:

Let:

E_i = total score of i-th examinee on m sampled items

I_j = total of scores of n sampled examinees on j-th item

X_{ij} = score of i-th examinee on j-th item

Define:

$$V_E = \frac{1}{m} S^2_E = \frac{1}{m(n-1)} \left[\sum_{i=1}^{n} E_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} E_i \right)^2 \right]$$

$$V_I = \frac{1}{n} S^2_I = \frac{1}{n(m-1)} \left[\sum_{j=1}^{m} I_j^2 - \frac{1}{m} \left(\sum_{j=1}^{m} I_j \right)^2 \right]$$

$$V_{EI} = \frac{1}{(n-1)(m-1)} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}^2 - \frac{1}{m} \sum_{i=1}^{n} E_i^2 - \frac{1}{n} \sum_{j=1}^{m} I_j^2 + \frac{1}{nm} \left(\sum_{i=1}^{n} E_i \right)^2 \right]$$

Then:

$$\hat{\sigma}^2_E = \frac{N-1}{N} \left[\frac{V_E - (1-m/M)\, V_{EI}}{m} \right]$$

$$\hat{\sigma}^2_I = \frac{M-1}{M} \left[\frac{V_I - (1-n/N)\, V_{EI}}{n} \right]$$

$$\hat{\sigma}^2_{EI} = \frac{(N-1)(M-1)}{NM} V_{EI}$$

Table 2

ESTIMATORS OF UNIVERSE SCORES AND THEIR STANDARD ERRORS

ESTIMATOR	STANDARD ERROR
(11) f = number of sampled question-examinee contacts that result in correct answer (p = f/nm) = (estimator)	$S(p) = \left\{ \frac{1}{mn(N-1)} \left[n(N-1)\hat{\sigma}^2_I + m(N-n)\hat{\sigma}^2_E + (N-n)\hat{\sigma}^2_{EI.} \right] \right\}^{1/2}$, where variance component estimators are defined as follows: Let: E_i, I_j, V_E, V_I, and V_{EI} be defined as in Equation (10). Then: $\hat{\sigma}^2_E = \frac{N}{N-1} \left[\frac{V_E - V_{EI}}{m} \right]$; $\hat{\sigma}^2_I = \left[\frac{V_I - (1-n/N)V_{EI}}{n} \right]$; $\hat{\sigma}^2_{EI} = \frac{N-1}{N} V_{EI}$

Table 2

ESTIMATORS OF UNIVERSE SCORES AND THEIR STANDARD ERRORS

ESTIMATOR	STANDARD ERROR
(12) f_i = number of sampled question-examinee contacts that result in correct answer for examinees in i-th stratum. $p_i = \left(\frac{f_i}{n_i m}\right)$ $N = \sum_{i=1}^{K} N_i$ $W_i = \frac{N_i}{N}$ $(\bar{p} = \sum_{i=1}^{K} W_i p_i) = (\text{estimator})$	$S(p) = \left[\sum_{i=1}^{K} W_i^2 S^2(p_i)\right]^{1/2}$, where $S^2(\dot{p}_i) = \frac{1}{mn_i(M-1)(N_i-1)}\Big[(M-m)n_i(N_i-1)\hat{\sigma}^2_{I_i} + m(M-1)(N_i-n_i)\hat{\sigma}^2_{E_i} + (M-m)(N_i-n_i)\hat{\sigma}^2_{(EI)_i}\Big]$,

and the variance components for the i-th stratum are defined as follows:

Let:

E_{ij} = total score of j-th examinee in i-th stratum on m sampled items.

I_{ik} = total of scores of n_i sampled examinees in i-th stratum on k-th item

X_{ijk} = score of j-th examinee in i-th stratum on k-th item

Define:

$$V_{E_i} = \frac{1}{m} S_{E_i}^2 = \frac{1}{m(n_i-1)}\left[\sum_{j=1}^{n_i} E_{ij}^2 - \frac{1}{n_i}\left(\sum_{j=1}^{n_i} E_{ij}\right)^2\right]$$

$$V_{I_i} = \frac{1}{n_i} S_{I_i}^2 = \frac{1}{n_i(m-1)}\left[\sum_{k=1}^{m} I_{ik}^2 - \frac{1}{m}\left(\sum_{k=1}^{m} I_{ik}\right)^2\right]$$

$$V_{(EI)_i} = \frac{1}{(n_i-1)(m-1)}\left[\sum_{j=1}^{n_i}\sum_{k=1}^{m} X_{ijk}^2 - \frac{1}{m}\sum_{j=1}^{n_i} E_{ij}^2 - \frac{1}{n_i}\sum_{k=1}^{m} I_{ik}^2 + \frac{1}{n_i m}\left(\sum_{j=1}^{n_i} E_{ij}\right)^2\right]$$

Table 2

ESTIMATORS OF UNIVERSE SCORES AND THEIR STANDARD ERRORS

ESTIMATOR	STANDARD ERROR
	Then: $\hat{\sigma}^2_{E_i} = \frac{N_i - 1}{N_i}\left[\frac{V_{E_i} - (1-m/M)V_{(EI)_i}}{m}\right]$ $\hat{\sigma}^2_{I_i} = \frac{M-1}{M}\left[\frac{V_{I_i} - (1-\frac{n_i}{N_i})V_{(EI)_i}}{n_i}\right]$ $\hat{\sigma}^2_{(EI)_i} = \frac{(N_i - 1)(M-1)}{N_i M} V_{(EI)_i}$
(13) f_i = number of sampled question-examinee contacts that result in correct answer for [illegible] examinees in the i-th stratum. $p_i = \left(\frac{f_i}{n_i m}\right)$ $N = \sum_{i=1}^{K} N_i$ $W_i = \frac{N_i}{N}$ $(p = \sum_{i=1}^{K} W_i p_i) = $ (estimator)	$S(p) = \left[\sum_{i=1}^{K} W_i^2 S^2(p_i)\right]^{1/2}$ $S(p_i) = \frac{1}{mn_i(N_i-1)}\left[n_i(N_i-1)\hat{\sigma}^2_{I_i} + m(N_i-n_i)\hat{\sigma}^2_{E_i} + (N_i-n_i)\hat{\sigma}^2_{(EI)_i}\right]$ where variance component estimators are defined as follows: Let: E_{ij}, I_{ik}, X_{ijk}, V_{E_i}, V_{I_i}, and $V_{(EI)_i}$ be defined as in Equation (12). Then: $\hat{\sigma}^2_{E_i} = \frac{N_i - 1}{N_i}\left[\frac{V_{E_i} - V_{(EI)_i}}{m}\right]$ $\hat{\sigma}^2_{I_i} = \left[\frac{V_{I_i} - (1-\frac{n_i}{N_i}) V_{(EI)_i}}{n_i}\right]$ $\hat{\sigma}^2_{(EI)_i} = \frac{N_i - 1}{N_i} V_{(EI)_i}$